Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation*

Tilmann Gneiting, Anton H. Westveld III, Adrian E. Raftery
and Tom Goldman

Department of Statistics
University of Washington, Seattle, Washington

---

*Corresponding author address: Tilmann Gneiting, Department of Statistics, University of Washington, Box 354320, Seattle, WA 98195-4322, e-mail: tilmann@stat.washington.edu.

Report Documentation Page

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **05 MAY 2004**	2. REPORT TYPE	3. DATES COVERED **00-05-2004 to 00-05-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**35**	

Abstract

Ensemble prediction systems typically show positive spread-error correlation, but they are subject to forecast bias and underdispersion, and therefore uncalibrated. This work proposes the use of ensemble model output statistics (EMOS), an easy to implement post-processing technique that addresses both forecast bias and underdispersion and takes account of the spread-skill relationship. The technique is based on multiple linear regression and akin to the superensemble approach that has traditionally been used for deterministic-style forecasts. The EMOS technique yields probabilistic forecasts that take the form of Gaussian predictive probability density functions (PDFs) for continuous weather variables, and can be applied to gridded model output. The EMOS predictive mean is an optimal, bias-corrected weighted average of the ensemble member forecasts, with coefficients that are constrained to be nonnegative and associated with the member model skill. The EMOS predictive mean provides a highly accurate deterministic-style forecast. The EMOS predictive variance is a linear function of the ensemble spread. For fitting the EMOS coefficients, the method of minimum CRPS estimation is introduced. The minimum CRPS estimator finds the coefficient values that optimize the continuous ranked probability score (CRPS) for the training data. The EMOS technique was applied to 48-hour forecasts of sea level pressure and surface temperature over the North American Pacific Northwest in Spring 2000, using the University of Washington mesoscale ensemble. When compared to the bias-corrected ensemble, deterministic-style EMOS forecasts of sea level pressure had root-mean-square error 9% less and mean absolute error 8% less. The EMOS predictive PDFs were much better calibrated than the raw ensemble or the bias-corrected ensemble, and they were sharp in that prediction intervals were considerably shorter on average than those obtained from climatological forecasts. Perhaps surprisingly, the EMOS ensemble was frequently sharper than the raw ensemble. When compared to the bias-corrected ensemble, EMOS improved the continuous ranked probability score by 16%. It also improved the ignorance score by 3.7, corresponding to the predictive PDF at the verifying observation being greater by a factor of 40.

1 Introduction

During the past decade, the use of forecast ensembles for assessing the uncertainty of numerical weather predictions has become routine. Three operational methods for the generation of synoptic-scale ensembles have been developed: The breeding growing modes method used by the US National Centers for Environmental Prediction (Toth and Kalnay 1993), the singular vector method used by the European Centre for Medium-Range Weather Forecasts (Molteni et al. 1996), and the perturbed observations method used by the Canadian Meteorological Centre (Houtekamer et al. 1996). More recently, mesoscale short-range ensembles have been developed, such as the University of Washington ensemble system over the North American Pacific Northwest (Grimit and Mass 2002; Eckel 2003). The ability of ensemble systems to improve deterministic-style forecasts and to predict forecast skill has been convincingly established. Statistically significant spread-error correlations suggest that ensemble variance and related measures of ensemble spread are skillful indicators of the accuracy of the ensemble mean forecast.

Case studies in probabilistic weather forecasting have typically focused on the prediction of categorical events. Ensembles also allow for probabilistic forecasts of continuous weather variables, such as air pressure and temperature, that are expressed in terms of predictive probability density functions (PDFs) or predictive cumulative distribution functions (CDFs). Due to the limited size of current ensemble systems, which typically consist of five to fifty ensemble member forecasts, raw ensemble output does not provide predictive PDFs, and some form of post-processing is required (Richardson 2001). However, various challenges in the statistical post-processing of ensemble output have been described. Systematic biases are substantial in current modeling systems (Atger 2003; Mass 2003) and might disguise probabilistic forecast skill. Furthermore, forecast ensembles are typically underdispersive (Hamill and Colucci 1997; Eckel and Walters 1998).

In this paper, we propose the use of ensemble model output statistics (EMOS), an easy to implement statistical post-processing technique that addresses the aforementioned issues. Our method is a variant of multiple linear regression or model output statistics (MOS) techniques that have traditionally been used for deterministic-style and probability of precipitation forecasts (Glahn and Lowry 1972; Wilks 1995). Specifically,

suppose that $X_1, \ldots, X_m$ denotes an ensemble of individually distinguishable forecasts for a univariate weather quantity Y. A multiple linear regression equation for Y in terms of the ensemble member forecasts can be written as

$$Y = a + b_1 X_1 + \cdots + b_m X_m + \varepsilon \tag{1}$$

where a and $b_1, \ldots, b_m$ are regression coefficients, and where ϵ is an error term that averages to zero. Regression approaches of this type have been shown to improve the deterministic-style forecast accuracy of synoptic weather and seasonal climate ensembles (Krishnamurti et al. 1999, 2000; Kharin and Zwiers 2002), and the associated forecast systems have been referred to as superensembles.

The use of regression techniques for probabilistic forecasting has not received much attention in the literature, with the exception of forecasts of binary events (Glahn and Lowry 1972; Stefanova and Krishnamurti 2002). In this work, we obtain full predictive PDFs and CDFs from ensemble forecasts of a continuous weather variable. Standard regression theory suggests a straightforward way of constructing predictive PDFs and CDFs from a regression equation, by taking them to be Gaussian with predictive mean equal to the regression estimate, and predictive variance equal to the mean squared prediction error for the training data. This approach corrects for model biases and takes account of underdispersion. However, the resulting assessment of uncertainty is static, in that the predictive variance is independent of the ensemble spread, thereby negating the spread-skill relationship (Whitaker and Loughe 1998). Hence, we model the variance of the error term in the multiple linear regression equation (1) as a linear function of the ensemble spread, that is,

$$\mathrm{Var}(\varepsilon) = c + dS^2 \tag{2}$$

where S^2 is the ensemble variance, and where c and d are nonnegative coefficients. Combining (1) and (2) yields the Gaussian predictive distribution

$$\mathcal{N}\left(a + b_1 X_1 + \cdots + b_m X_m, \; c + dS^2\right)$$

whose mean equals the regression estimate and forms a bias-corrected weighted average of the ensemble member forecasts, and whose variance depends linearly on the ensemble spread. Negative regression weights can, and frequently do, occur in this type of formulation, as in Tables 2, 4, 5, and 6 of van den Dool and Rukhovets (1994). This is an

4

Table 1: Phase I of the University of Washington mesoscale short-range ensemble, January–June 2000. Initial conditions (ICs) and lateral boundary conditions (LBCs) were obtained from the Aviation Model (AVN), the Nested Grid Model (NGM) Regional Data Assimilation System, and the Eta Data Assimilation System, all run by the US National Centers for Environmental Prediction (NCEP), the Global Environmental Multiscale (GEM) analysis run by the Canadian Meteorological Centre (CMC), and the US Navy Operational Global Atmospheric Prediction System (NOGAPS) analysis run by Fleet Numerical Meteorology and Oceanography Center (FNMOC). See Grimit and Mass (2002) for details.

No.	Ensemble Member	IC/LBC Source
1	AVN-MM5	NCEP
2	GEM-MM5	CMC
3	ETA-MM5	NCEP
4	NGM-MM5	NCEP
5	NOGAPS-MM5	FNMOC

artifact caused by the collinearity of the ensemble member forecasts, and the negative weights seem hard to interpret. They imply, all else equal, that sea level pressure, say, is predicted to be lower when the forecast with the negative weight is higher. To address this issue, we estimate the statistical model under the constraint that the coefficients $b_1, \ldots, b_m$, as well as c and d, are nonnegative. We refer to the resulting predictive PDFs and CDFs as ensemble model output statistics or EMOS forecasts.

We applied the EMOS technique to the University of Washington mesoscale short-range ensemble described by Grimit and Mass (2002). Briefly, this is a multi-analysis, single-model (MM5) ensemble driven by initial conditions and lateral boundary conditions obtained from major operational weather centers worldwide. Table 1 provides an overview of the Phase I University of Washington ensemble system. Figure 1 illustrates the spread-skill relationship for sea level pressure forecasts, using the same period January – June 2000 on which the study of Grimit and Mass (2002) was based. The ensemble spread provides useful information about the error of the ensemble mean forecast. Figure 2 gives an example of a 48-hour EMOS forecast of sea level pressure. This forecast was initialized 0000 UTC 25 May 2000 and was valid at Hope Airport, British Columbia. Both the EMOS predictive PDF and the EMOS predictive CDF are shown.

Spread–Skill Relationship for Sea–Level Pressure Forecasts
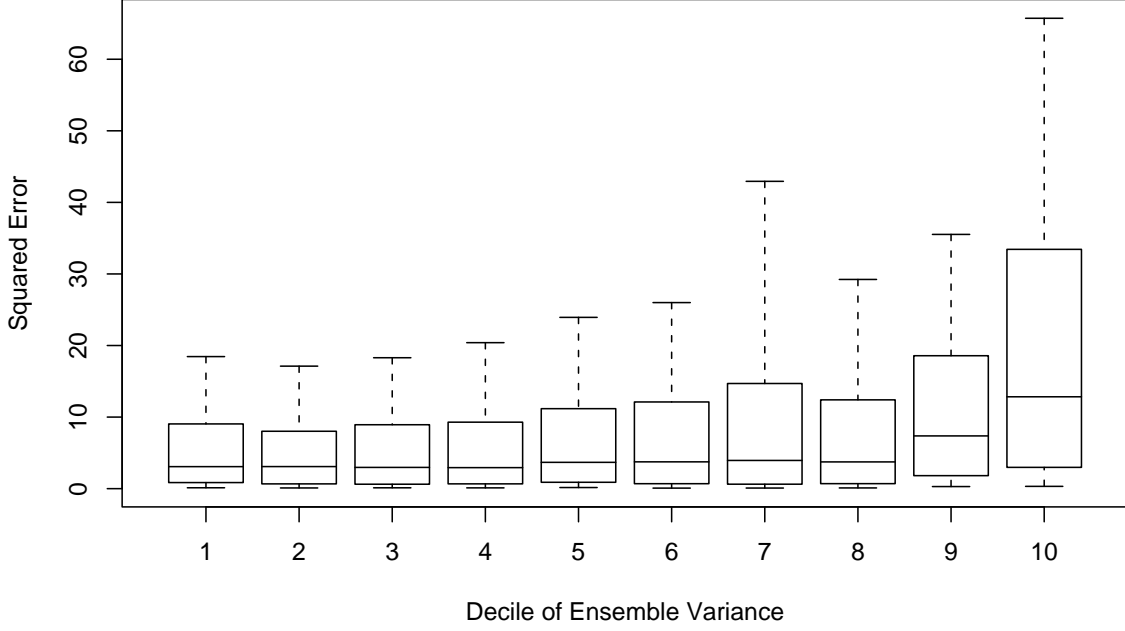

Figure 1: Spread-skill relationship for ensemble mean forecasts of sea level pressure over the Pacific Northwest, January – June 2000. For each decile of the ensemble variance, the boxplots show the 10%, 25%, 50%, 75%, and 90% percentile of the squared forecast error. The correlation coefficient between the ensemble spread and the squared forecast error was 0.33 for individual forecasts, and 0.52 for daily averages aggregated across the Pacific Northwest.

The construction of prediction intervals from the predictive CDF, say F, is straightforward. For instance, $F(\frac{1}{6})$ and $F(\frac{5}{6})$ form the lower and upper endpoint of the $66\frac{2}{3}\%$ central prediction interval, respectively. In the Hope Airport example, and using the millibar as unit, this interval was [1008.3, 1013.0]. The ensemble range of the University of Washington ensemble was [1003.7, 1016.8]. For a five-member ensemble, this is also a nominal $66\frac{2}{3}\%$ prediction interval, but is much wider. Perhaps surprisingly, this situation – EMOS prediction intervals that are shorter than their ensemble counterparts – was not uncommon. In our case study this occurred in about 27% of the sea level pressure forecasts.

The paper is organized as follows. In Section 2 we describe the EMOS technique in detail, and we explain how we go about verifying probabilistic forecasts. In assessing
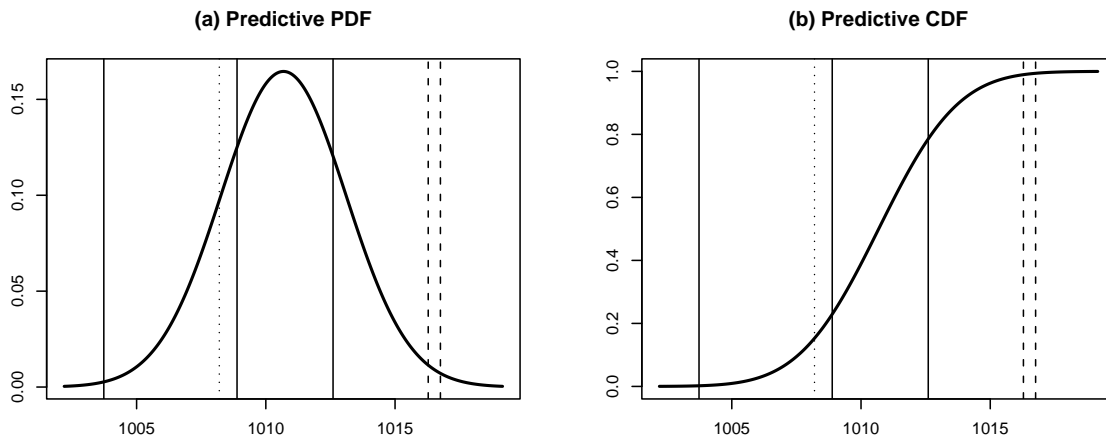
Figure 2: EMOS 48-hour forecast of sea level pressure at Hope Airport, British Columbia, initialized 0000 UTC 25 May 2000: (a) Predictive PDF. (b) Predictive CDF. Also shown are the five ensemble member forecasts (solid and broken lines) and the verifying observation (dotted line). The ETA-MM5 and NGM-MM5 forecasts (broken lines) had zero EMOS weight. The unit used is the millibar (mb).

forecast PDFs, we are guided by the principle that probabilistic forecasts strive to maximize sharpness subject to calibration (Gneiting et al. 2003). We apply diagnostic tools, such as the verification rank histogram and the probability integral transform (PIT) histogram, as well as scoring rules, among them the continuous ranked probability score (CRPS) and ignorance score. For estimating the EMOS coefficients, we introduce the novel approach of minimum CRPS estimation, which forms a particular case of minimum contrast estimation. Specifically, we find the coefficient values that minimize the continuous ranked probability score for the training data. For EMOS, this method gives better results than classical maximum likelihood estimation, which is nonrobust and favors overdispersive forecast PDFs.

Section 3 provides a case study of EMOS forecasts for sea level pressure and surface temperature in Spring 2000 over the Pacific Northwest, using the University of Washington ensemble. We explain how we find a suitable training period, and we describe and verify the EMOS forecasts. The EMOS forecast PDFs were much better calibrated than the raw ensemble or the bias-corrected ensemble, and the mean absolute error (MAE), root-mean-square error (RMSE), continuous ranked probability score (CRPS), and ig-

norance score (IGN) for the EMOS forecasts were consistently, and substantially, better than the corresponding quantities for the raw ensemble or the bias-corrected ensemble. The paper closes with a discussion in Section 4.

2   Methods

We now explain our approach to verifying probabilistic forecasts, and we describe the ensemble model output statistics (EMOS) technique in detail. For estimating the EMOS coefficients we use the novel approach of minimum CRPS estimation, which forms a special case of minimum contrast estimation (MCE). This method is best explained in terms of verification measures, so we describe these first.

2.1   Assessing sharpness and calibration

The goal of probabilistic forecasting is to maximize the sharpness of the forecast PDFs subject to calibration (Gneiting et al. 2003). Calibration refers to the statistical consistency between the forecast PDFs and the verifications, and is a joint property of the predictions and the verifications. Briefly, a forecast technique is calibrated if meteorological events declared to have probability p occur a proportion p of the time on average. Sharpness refers to the spread of the forecast PDFs and is a property of the predictions only. A forecast technique is sharp if prediction intervals are shorter on average than prediction intervals derived from naive methods, such as climatology or persistence. The more concentrated the forecast PDFs are, the sharper the forecast, and the sharper the better, subject to calibration.

The principal tool for assessing the calibration of ensemble forecasts is the verification rank histogram or Talagrand diagram (Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997; Hamill 2001). To obtain a verification rank histogram, find the rank of the verifying observation when pooled within the ordered ensemble values, and plot the histogram of the ranks.

The analogous tool for PDF forecasts is the probability integral transform (PIT) histogram. If F denotes the predictive CDF, the probability integral transform is simply the value $F(x)$ at the verification x, a number between 0 and 1. For the Hope Airport forecast in Figure 2(b), for instance, the PIT value was 0.15. Rosenblatt (1952) studied

the probability integral transform, and Dawid (1984) proposed its use in the assessment of probabilistic forecasts. The PIT histogram – that is, the histogram of the PIT values – is a commonly used tool in the econometric literature on probabilistic forecasting (see, for instance, Weigend and Shi 2000). Its interpretation is the same as that of the verification rank histogram: Calibrated probabilistic forecasts yield PIT histograms that are close to uniform, while underdispersive forecasts result in U-shaped PIT histograms.

How can ensembles and PDF forecasts be fairly compared? An ensemble provides a finite, typically small, number of values only, while PDF forecasts give continuous statements of uncertainty; so this seems difficult. There are two natural approaches to a fair comparison, using either the verification rank histogram or the PIT histogram. To obtain an m-member ensemble from a PDF forecast, take the CDF quantiles at levels $\frac{i}{m+1}$, for $i = 1, \ldots, m$. The verification rank histogram can then be formed in the usual way. To obtain a PIT histogram from an ensemble, fit a PDF to each ensemble forecast, as proposed by Déqué et al. (1994), Wilson et al. (1999), and Grimit and Mass (2004). The standard ensemble smoothing approach of Grimit and Mass (2004) fits a normal distribution with mean equal to the ensemble mean and variance equal to the ensemble variance. The PIT value is then computed on the basis of the fitted Gaussian CDF. Wilks (2002) proposed to smooth forecast ensembles by fitting mixtures of Gaussian distributions, an approach that allows for multimodal forecast PDFs. Multimodality may indeed be an issue for larger ensembles. For smaller ensembles, such as the University of Washington ensemble, standard ensemble smoothing using a single normal density suffices.

In addition to showing verification rank histograms and PIT histograms, we report the coverage of the $66\frac{2}{3}\%$ central prediction interval; we chose this interval, because the range of a five-member ensemble provides such. Finally, to assess sharpness, we consider the average width of the $66\frac{2}{3}\%$ prediction intervals. For a five-member ensemble, this is just the average ensemble range.

For Gaussian predictive PDFs, the average width of the $100 \times (1 - \alpha)\%$ prediction intervals is

$$2 z_{1-\frac{\alpha}{2}} \bar{s} \tag{3}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the $(1 - \frac{\alpha}{2})$ quantile of the normal distribution with mean 0 and variance 1, respectively, and where $\bar{s}$ stands for the average standard deviation of the

predictive PDFs. For instance, Table 2 below shows that the average width of the central $66\frac{2}{3}\%$ prediction intervals for MCE-EMOS forecasts of sea level pressure is 4.747. From (3) with $\alpha = \frac{1}{3}$ we find that $\bar{s} = 2.45$. Using again (3), the average width of the 50% and 90% central prediction intervals is 3.31 and 8.08, respectively.

2.2 Scoring rules

Scoring rules for the verification of deterministic-style forecasts are well-known and have been widely used in forecast assessment. If μ_i denotes a deterministic-style forecast and y_i is the verification, the mean absolute error (MAE) is defined as

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \mu_i|$$

where the sum is taken over the test data. A related error measure is the mean-square error (MSE), defined by

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_i)^2 \, .$$

The root-mean-square error (RMSE) is the square root of the MSE and has the advantage of being recorded in the same unit as the verifications.

We also consider two scoring rules for the assessment of predictive PDFs, the continuous ranked probability score (Unger 1985; Hersbach 2000; Gneiting and Raftery 2004), and the ignorance score (Good 1952; Roulston and Smith 2002). These scoring rules are attractive in that they address calibration as well as sharpness.

The continuous ranked probability score (CRPS) is the integral of the Brier scores at all possible threshold values t for the continuous predictand (Hersbach 2000; Toth et al. 2003, Section 7.5.2). Specifically, if F is the predictive CDF and y verifies, the continuous ranked probability score is defined as

$$\mathrm{crps}(F, y) = \int_{-\infty}^{\infty} \left(F(t) - H(t - y) \right)^2 \, \mathrm{d}t \tag{4}$$

where $H(t-y)$ denotes the Heaviside function and takes the value 0 when $t < y$ and the value 1 otherwise. Applications of the continuous ranked probability score have been hampered by a lack of closed form expressions for the associated integral. However,

when F is the CDF of a normal distribution with mean μ and variance σ^2, repeated partial integration in (4) shows that

$$\text{crps}(\mathcal{N}(\mu, \sigma^2), y) = \sigma \left(\frac{y - \mu}{\sigma} \left(2\,\Phi\left(\frac{y - \mu}{\sigma}\right) - 1 \right) + 2\,\varphi\left(\frac{y - \mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right) \tag{5}$$

where φ and Φ denote the PDF and the CDF, respectively, of the normal distribution with mean 0 and variance 1. A key difference between the ignorance score and the continuous ranked probability score is that (5) grows linearly in the normalized prediction error, $z = (y - \mu)/\sigma$, while (8) grows quadratically in z. Hence, the ignorance score assigns harsh penalties to particularly poor probabilistic forecasts, and can be exceedingly sensitive to outliers and extreme events (Weigend and Shi 2000; Gneiting and Raftery 2004). This will become apparent in Tables 5 and 7 below. Returning to the continuous ranked probability score, we note from (4) that the average score

$$\text{CRPS} = \frac{1}{n} \sum_{i=1}^{n} \text{crps}(F_i, y_i) \tag{6}$$

reduces to the MAE if each F_i is a deterministic-style forecast. For this and other reasons, the CRPS can be interpreted as a probabilistic version of the MAE.

The ignorance score is the negative of the logarithm of the predictive density f at the verifying value y, that is, for a single PDF forecast,

$$\text{ign}(f, y) = -\log f(y). \tag{7}$$

Roulston and Smith (2002) provide an interesting information theoretic perspective on the ignorance score. In the case of a normal predictive PDF with mean μ and variance σ^2, we have

$$\text{ign}(\mathcal{N}(\mu, \sigma^2), y) = \frac{1}{2} \ln(2\pi\sigma^2) + \frac{(y - \mu)^2}{2\sigma^2} \tag{8}$$

and the average ignorance is

$$\text{IGN} = \frac{1}{n} \sum_{i=1}^{n} \text{ign}(F_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \ln(2\pi\sigma_i^2) + \frac{(y_i - \mu_i)^2}{2\sigma_i^2} \right). \tag{9}$$

When interpreting improvements in the IGN score, it is absolute rather than relative changes that are relevant. An improvement of Δ in the IGN score corresponds to an increase in the predictive PDF at the verifying values by a factor of e^{Δ}.

11

Both CRPS and IGN are negatively oriented scores, in that a smaller value is better, and both scores are proper, meaning that they reward honest assessments. We report both scores, but in view of the lack of robustness of the ignorance score, we prefer the continuous ranked probability score. A more detailed discussion of scoring rules is given in Gneiting and Raftery (2004).

2.3 Ensemble model output statistics and minimum CRPS estimation

We now describe the ensemble model output statistics (EMOS) method. Suppose that $X_1, \ldots, X_m$ denotes an ensemble of forecasts for a univariate weather quantity Y, and let S^2 be the ensemble variance. The EMOS predictive PDF is that of the normal distribution

$$\mathcal{N}\left(a + b_1 X_1 + \cdots + b_m X_m,\ c + dS^2\right). \tag{10}$$

The EMOS predictive mean $a + b_1 X_1 + \cdots + b_m X_m$ is an optimal bias-corrected weighted average of the ensemble member forecasts. It provides a highly accurate deterministic-style forecast. The EMOS predictive variance $c + dS^2$ is a linear function of the ensemble spread. The regression coefficients $b_1, \ldots, b_m$ in (10) reflect the individual member model skill. Stefanova and Krishnamurti (2002) argue similarly in a superensemble context, but they do not constrain the regression coefficients $b_1, \ldots, b_m$ to be nonnegative. The variance coefficients c and d can be interpreted in terms of the ensemble spread and the skill of the ensemble mean forecast. All else equal, larger values of the coefficient d suggest a more pronounced spread-skill relationship. If spread and error are independent of each other, the coefficient d will be estimated as negligibly small. Hence, EMOS is robust, in the sense that it adapts to the presence as well as to the absence of significant spread-error correlation.

A classical technique for estimating the coefficients a, $b_1, \ldots, b_m$, c, and d from training data is maximum likelihood (Wilks 1995, Section 4.7). The likelihood function is defined as the probability of the training data given the coefficients, viewed as a function of the coefficients. In practice, it is more convenient to maximize the logarithm of the likelihood function, for reasons of both algebraic simplicity and numerical stability. The

log-likelihood function for the statistical model (10) is

$$\ell\left(a;\, b_1,\ldots,b_m;\, c;\, d\right) \tag{11}$$
$$= -\frac{1}{2}\left(k\log(2\pi) + \sum_{i=1}^{k}\frac{(Y_i - (a + b_1 X_{i1} + \cdots + b_m X_{im}))^2}{c + dS_i^2} + \sum_{i=1}^{k}\log\left(c + dS_i^2\right)\right)$$

where the sum is taken over the training data; here, $X_{i1},\ldots,X_{im}$ denote the ith ensemble forecast in the training set, S_i^2 denotes its variance, and Y_i denotes the ith verification, respectively. Strictly speaking, (11) is the log-likelihood function under the assumption of independence. Note that the log-likelihood (11) is essentially the negative of the ignorance score (9), but is applied to the training data rather than the test data. Hence, maximum likelihood estimation is equivalent to minimizing the ignorance score for the training data.

This observation suggests a general estimation strategy: Pick a scoring rule that is relevant to the problem at hand, express the score for the training data as a function of the coefficients, and optimize that function with respect to the coefficient values. We take scoring rules to be negatively oriented, so a smaller value is better, and we minimize the training score. For positively oriented scoring rules, we would maximize the training score. Such an approach is formally equivalent to minimum contrast estimation (MCE), a technique that has been studied in the theoretical statistics literature (Pfanzagl 1969, Birgé and Massart 1993). The minimum score approach can also be interpreted within the framework of robust M-estimation (Huber 1964; Huber 1981, Section 3.2) and forms a special case thereof, in that the function to be optimized derives from a strictly proper scoring rule (Gneiting and Raftery 2004). A more detailed methodological and theoretical discussion is beyond the scope of this paper. However, we compared EMOS PDF forecasts estimated by MCE with the continuous ranked probability score, as described below, to EMOS PDF forecasts estimated by maximum likelihood. The former performed clearly better: the predictive PDFs were better calibrated, and they were sharper. This comparison is summarized in Table 2. As a rule of thumb, it seems that predictive PDFs estimated by maximum likelihood tend to be overdispersive, resulting in unnecessarily wide prediction intervals that have higher than nominal coverage, and in inverted U-shaped PIT histograms. This latter shape is also seen in Figures 4 and 5 of Weigend and Shi (2000), who estimate predictive densities by the maximum likelihood method in the form of the EM algorithm.

Table 2: Comparison of EMOS predictive PDFs obtained by maximum likelihood estimation (MLE-EMOS) and minimum CRPS estimation (MCE-EMOS), respectively. The results are for the test data, region, and 40-day sliding training period described in Section 3.

	Score		Score		$66\frac{2}{3}\%$ Prediction Interval	
	MAE	RMSE	CRPS	IGN	Coverage	Average Width
Sea level pressure						
MLE-EMOS	1.958	2.496	1.391	2.323	69.41	4.931
MCE-EMOS	1.953	2.487	1.389	2.326	67.61	4.747
Surface temperature						
MLE-EMOS	2.239	2.914	1.614	2.490	72.69	5.909
MCE-EMOS	2.230	2.906	1.606	2.488	68.57	5.411

We argued in Section 2.2 that the continuous ranked probability score (CRPS) is a more robust and therefore more appropriate scoring rule than the ignorance score. This suggests the use of the continuous ranked probability score in minimum contrast estimation; and this might be called minimum CRPS estimation. The minimum CRPS estimator finds the coefficients a, $b_1, \ldots, b_m$, c, and d in the statistical model (10) that minimize the CRPS value for the training data. Using (5) and (6), we express the training CRPS as an analytic function of the coefficients, namely

$$\Gamma\left(a;\, b_1, \ldots, b_m;\, c;\, d\right) = \frac{1}{k} \sum_{i=1}^{k} \left(c + dS_i^2\right) \left(Z_i \left(2\,\Phi(Z_i) - 1\right) + 2\,\varphi(Z_i) - \frac{1}{\sqrt{\pi}}\right) \quad (12)$$

where

$$Z_i = \frac{(Y_i - (a + b_1 X_{i1} + \cdots + b_m X_{im}))^2}{c + dS_i^2}$$

is a standardized forecast error, and where φ and Φ denote the PDF and CDF, respectively, of a normal distribution with mean 0 and variance 1. We find the coefficient values that minimize (12) numerically, using the Broyden-Fletcher-Goldfarb-Shanno algorithm (Press et al. 1992, Section 10.7) as implemented in the R language and environment (www.cran.r-project.org/). The optimization algorithm requires initial values, and one way of specifying them is by least squares estimation for the standard multiple linear

14

regression model (1). In the sliding window implementation of Section 3, we use the previously estimated EMOS coefficients as initial values in the subsequent optimization problem.

How do we constrain the coefficients $b_1, \ldots, b_m$, c, and d to be nonnegative? The nonnegativity of the variance term c is not an issue. To enforce the nonnegativity of d, we set $d = \delta^2$ and optimize over δ; this turned out to be numerically stable. To enforce nonnegative regression coefficients, we proceed stepwise. We first find the unconstrained minimum of the CRPS value (12). If all estimated regression coefficients are nonnegative, the EMOS model is complete. If one or more of the regression coefficients are negative, we set these to zero, and we minimize the CRPS value (12) under that constraint. We also re-compute the ensemble variance, using only the ensemble members that remain in the regression equation, and we subsequently use the re-computed ensemble spread. This procedure is iterated until all estimated regression coefficients are nonnegative.

Table 3 illustrates this algorithm for predictions on 25 May 2000, the day on which the Hope Airport forecast in Figure 2 was issued. The initial, unconstrained minimization of the CRPS value (12) uses the EMOS coefficients from the previous fit as initial values. This results in a negative coefficient for the third ensemble member model, the ETA-MM5 forecast. We set this coefficient to zero and proceed with the constrained minimization, resulting in a negative weight for the NGM-MM5 forecast. The final EMOS equation uses only one of the three ensemble members initialized with NCEP models, namely the AVN-MM5 forecast, along with the GEM-MM5 forecast, and the NOGAPS-MM5 forecast. The EMOS weights reflect the relative skill of the ensemble member models during the 40-day training period. Indeed, the bias-corrected AVN-MM5, GEM-MM5, and NOGAPS-MM5 forecasts had a smaller training RMSE than the bias-corrected ETA-MM5 and NGM-MM5 forecasts. The AVN-MM5, ETA-MM5, and NGM-MM5 forecasts were initialized by NCEP models, and they were highly collinear. For the training period, the pair correlation coefficients within this group ranged from 0.93 to 0.97. EMOS picked and retained the most skillful among the three collinear forecasts. The correlation coefficients between NCEP- and non NCEP-initialized member model forecasts were also high, but they reached at most 0.92. The estimated variance coefficient d turned out to be negligibly small, thereby indicating a weak spread-skill relationship during the training period. Indeed, the correlation coefficient between the

Table 3: Minimum CRPS estimation of the EMOS coefficients for the Hope Airport forecast PDF in Figure 2. The regression coefficients $b_1, \ldots, b_5$ correspond to the AVN-MM5, GEM-MM5, ETA-MM5, NGM-MM5, and NOGAPS-MM5 forecast, respectively.

	a	b_1	b_2	b_3	b_4	b_5	c	d
Initial values	126.31	0.31	0.34	0.00	0.00	0.24	6.22	0.00
First stage	131.27	0.37	0.36	-0.17	0.02	0.29	5.77	0.00
Second stage	143.23	0.35	0.33	—	-0.09	0.28	5.81	0.00
EMOS	130.34	0.31	0.31	—	—	0.25	5.88	0.00

ensemble spread and the squared error of the ensemble mean forecast was only 0.11 for the 40-day training period that we used, as compared to 0.33 for the entire period, January–June 2000. The resulting EMOS predictive PDF is Gaussian, and is straightforward to simulate from. An alternative, and likely preferable, way of forming an m-member ensemble from the predictive PDF is by taking the forecast quantiles at level $\frac{i}{m+1}$, for $i = 1, \ldots, m$, respectively. In this way, ensembles of any size can be obtained, and in this sense, EMOS can be viewed as a dressing method (Roulston and Smith 2002).

It is worth pointing out that the EMOS model (10) can be estimated under further constraints. The general formulation requires that the ensemble members come from individually distinguishable sources. This is true for the University of Washington ensemble, a multi-analysis, mesoscale, short-range ensemble, and also for poor person's and multi-model ensembles. If the linear regression is based on the ensemble mean only, which constrains the regression coefficients $b_1 = \cdots = b_m$ in (10) to be equal, EMOS can be applied to essentially all ensemble systems, including perturbed observations, singular vector, and bred ensembles. Jewson et al. (2003) applied such an approach to the synoptic ECMWF ensemble, using maximum likelihood estimation. However, they did not report out of sample forecasts, and consequently neither verification scores nor rank histograms. For the University of Washington ensemble, the general formulation seems preferable. In the situation of Table 2, constraining the regression coefficients in (10) to be equal, that is, using the ensemble mean only, results in MAE, RMSE, and CRPS scores up to 7% worse, as compared to the full formulation.

3 Results for the University of Washington ensemble over the Pacific Northwest

We now give the results of applying EMOS to 48-hour forecasts of sea level pressure and surface temperature over the northwestern United States and British Columbia, using Phase I of the University of Washington ensemble described by Grimit and Mass (2002). The University of Washington ensemble system is a mesoscale, short-range ensemble based on the Fifth-generation Pennsylvania State University – National Center for Atmospheric Research Mesoscale Model (PSU-NCAR MM5) and forms an integral part of the Pacific Northwest regional environmental prediction effort (Mass et al. 2003). The ensemble used a 0000 UTC cycle and was in operation on 102 days between 12 January 2000 and 30 June 2000; it is described in Table 1. During this period, there were 16,015 and 56,489 verifying observations of sea level pressure and surface temperature, respectively. Model forecast data at the four grid points surrounding each observation were bilinearly interpolated to the observation site (Grimit and Mass 2002). When we talk of a 40-day training period, say, we refer to the most recent 40 days for which ensemble output and verifying observations were available. In terms of calendar days, this period typically corresponds to more than 40 days.

3.1 Length of training period

What training period should be used for estimating the EMOS regression coefficients and variance parameters? There is a trade-off here. Shorter training periods allow to adapt rapidly to seasonally varying model biases, changes in the performance of the ensemble member models, and changes in environmental conditions. On the other hand, longer training periods reduce the statistical variability in the estimation of the EMOS coefficients. We considered training periods of 19, 20, ..., 62 days for forecasts of sea level pressure. For comparability, the same test set was used in assessing all the training periods, that is, the first 63 days on which the ensemble was operating were not included in the test data. The unit used for the sea level pressure forecasts is the millibar (mb).

The results of this experiment are summarized in Figure 3. Figures 3(a) and 3(b) show the mean absolute error (MAE) and root-mean-square error (RMSE) of the deterministic-style EMOS forecasts, respectively. These decrease sharply for training periods
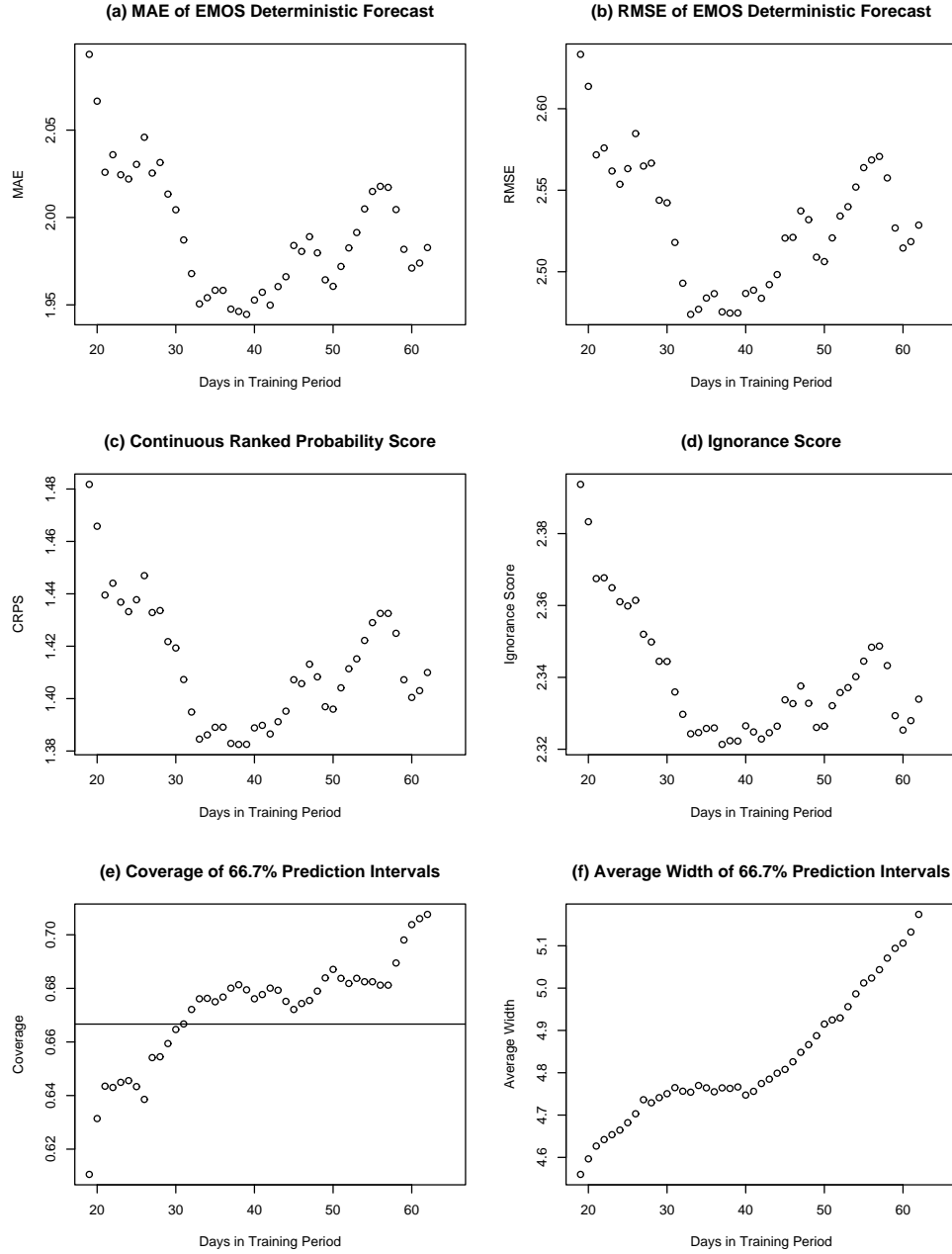
Figure 3: Comparison of training period lengths for forecasts of sea level pressure over the Pacific Northwest: (a) MAE of EMOS deterministic-style forecasts. (b) RMSE of EMOS deterministic-style forecasts. (c) Continuous ranked probability score. (d) Ignorance score. (e) Coverage of $66\frac{2}{3}\%$ prediction intervals. (f) Average width of $66\frac{2}{3}\%$ prediction intervals.

less than 30 days, stay about constant for training periods between 30 and 45 days, and increase thereafter. Figures 3(c) and 3(d) show the continuous ranked probability score (CRPS) and the ignorance score (IGN). The patterns are similar to those for the MAE and the RMSE. The coverage of EMOS $66\frac{2}{3}\%$ prediction intervals is shown in Figure 3(e). Training periods under 30 days seem to result in underdispersive PDFs, but training periods between 30 and 60 days show close to nominal coverage. Figure 3(f) shows the average widths of the $66\frac{2}{3}\%$ prediction intervals. The average width increases with the length of the training period, but is about constant for training periods between 30 and 40 days.

To summarize these results, there appear to be substantial gains in increasing the training period beyond 30 days. As the training period increases beyond 45 days, the skill of the probabilistic forecasts declines slowly but steadily, presumably as a result of seasonally varying model biases. In view of our goal of maximizing sharpness subject to calibration, we chose a 40-day training period. This worked well for temperature forecasts, too. However, distinct training periods might work best for distinct variables, forecast horizons, time periods, and regions. Ideally, we would include training data from previous years to address seasonal effects. Further research in this direction is desirable as multi-year runs of stable mesoscale ensembles become available.

3.2   Sea level pressure forecasts

We now give the results for EMOS forecasts of sea level pressure, using a 40-day sliding training period and the same test set that was used to compare the different training periods. We also summarize the results for the bias-corrected ensemble member forecasts and for a climatological forecast. The bias-corrected ensemble member forecasts were obtained by simple linear regression, estimated on the same 40-day sliding training period. The deterministic-style climatological forecast was the average sea level pressure among the verifying observations in the training period, and the climatological predictive PDF was obtained by fitting a normal PDF to the training data.

Figure 4 shows the estimates of the EMOS coefficients, as they evolve over the test period. The estimated intercept in the multiple linear regression equation is shown in Figure 4(a). Figures 4(b), (c), (d), (e), and (f) show the EMOS weights for the five ensemble member models, respectively. The weights for the AVN-MM5, CMC-MM5,
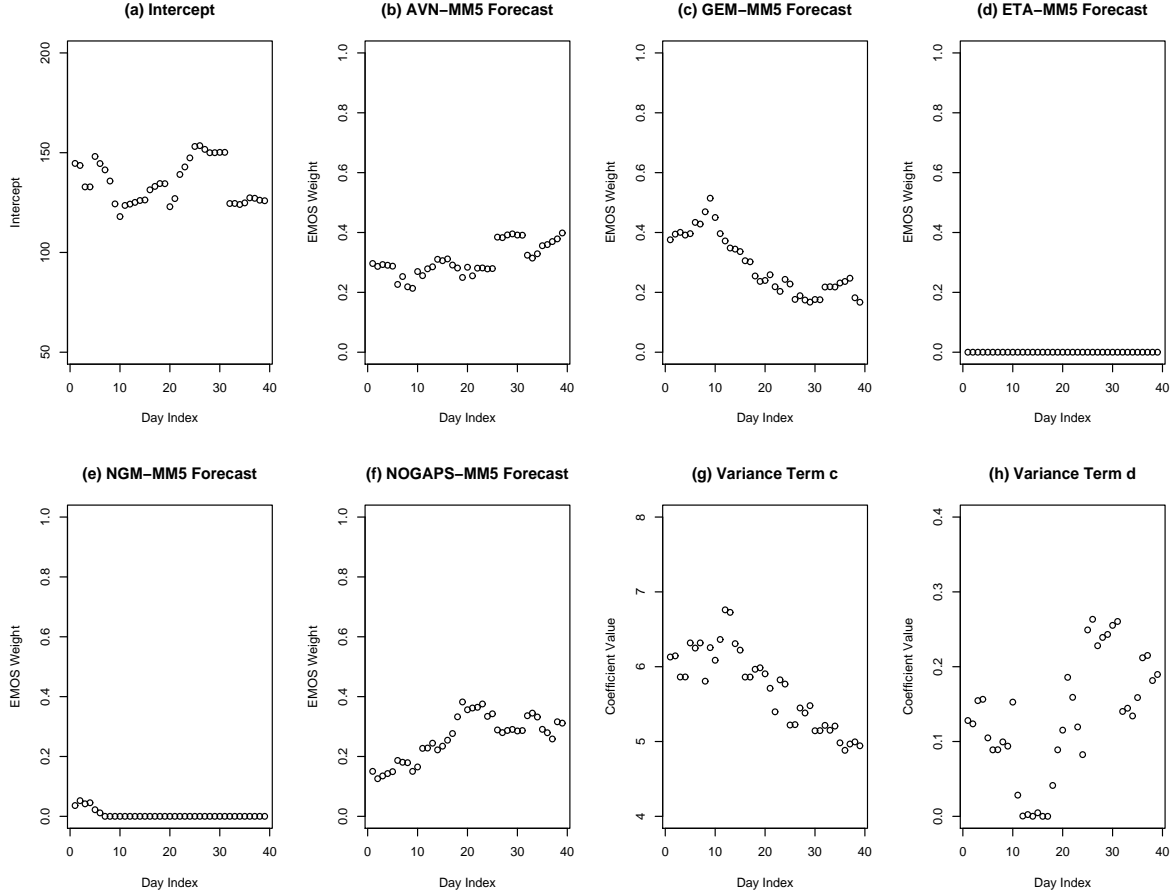
Figure 4: Coefficient estimates for EMOS forecasts of sea level pressure over the Pacific Northwest. (a) Intercept. (b), (c), (d), (e), and (f): Member model weights. (g) and (h): Variance terms c and d.

and NOGAPS-MM5 forecasts were consistently substantial, and the weights for the ETA-MM5 and NGM-MM5 forecasts were consistently negligible or zero. This is a collinearity effect, very much like the effect described in Section 2.3. EMOS retains only one of the three highly collinear ensemble member models that were initialized by NCEP analyses and picks the most skillful of them, namely the AVN-MM5 forecast. Figures 4(f) and 4(g) show the estimated variance coefficients c and d, respectively. The estimates of c decreased during the test period, thereby indicating improved ensemble skill or improved atmospheric predictability, or both. The values of d were small but mostly nonzero. The increase toward the end of the test period suggests a strengthening

of the spread-skill relationship.

Table 4 provides summary measures of deterministic-style forecast accuracy. Among the raw ensemble member models, the AVN-MM5 forecast performed best. Bias correction resulted in a reduction of the RMSE for the ensemble member model forecasts between 4% and 17%. The improvement was most pronounced for the NOGAPS-MM5 forecast. The AVN-MM5, CMC-MM5, and NOGAPS-MM5 forecasts were more accurate than the ETA-MM5 and NGM-MM5 forecasts. This confirms our interpretation of the EMOS weights in terms of the individual member model skill. The ensemble mean forecast performed considerably better than any of the ensemble member models. However, the bias-corrected AVN-MM5 forecast and the bias-corrected NOGAPS-MM5 forecast were more accurate than the mean of the bias-corrected ensemble. The deterministic-style EMOS forecast clearly performed best and had RMSE 9% and 8% less when compared to the mean of the raw ensemble and to the mean of the bias-corrected ensemble, respectively. The results in terms of the mean absolute error (MAE) were similar.

Table 5 turns to summary measures of probabilistic forecast skill. The climatological predictive PDFs showed the correct coverage, but they were too spread out to be competitive. The bias-corrected ensemble shows reduced ensemble spread, but is even more underdispersive than the raw ensemble. The EMOS prediction intervals show accurate coverage. The continuous ranked probability score (CRPS) and the ignorance score (IGN) were computed as described in Section 2.2, using standard ensemble smoothing for the raw and bias-corrected ensemble, respectively. The CRPS score can also be computed directly, by using the empirical ensemble CDF, which takes the values $0, \frac{1}{5}, \ldots, \frac{4}{5}, 1$, with jumps at the ensemble member forecasts. This gave somewhat higher CRPS values of 1.69 and 1.72 for the raw ensemble and for the bias-corrected ensemble, respectively. The EMOS predictive PDFs had by far the best scores among the different forecasts. When compared to the bias-corrected ensemble, EMOS reduced the CRPS score by 16%. EMOS reduced the IGN score by 3.68 points, indicating that the predictive PDF of verifying observations increased by a factor of 40. The EMOS prediction intervals were not much wider than prediction intervals obtained from the raw ensemble. A more detailed analysis shows, perhaps surprisingly, that in 27% of the forecasts the EMOS $66\frac{2}{3}\%$ prediction interval was shorter than the range of the five-member raw

Table 4: Comparison of deterministic-style forecasts of sea level pressure over the Pacific Northwest. The climatological, bias-corrected, and EMOS forecasts were trained on a sliding 40-day period.

	MAE	RMSE
Climatological forecast	4.72	5.83
AVN-MM5	2.20	2.90
GEM-MM5	2.35	3.00
ETA-MM5	2.50	3.25
NGM-MM5	2.70	3.40
NOGAPS-MM5	2.50	3.21
AVN-MM5 bias-corrected	2.10	2.68
GEM-MM5 bias-corrected	2.24	2.88
ETA-MM5 bias-corrected	2.37	3.14
NGM-MM5 bias-corrected	2.48	3.23
NOGAPS-MM5 bias-corrected	2.10	2.66
Mean of raw ensemble	2.11	2.73
Mean of bias-corrected ensemble	2.08	2.69
EMOS forecast	1.95	2.49

ensemble. In 9% of the forecasts, the EMOS $66\frac{2}{3}\%$ prediction interval was shorter than the range of the bias-corrected ensemble.

The verification rank histograms for the raw ensemble, the bias-corrected ensemble, and the EMOS ensemble are shown in Figure 5. The EMOS ensemble was much better calibrated than the raw ensemble or the bias-corrected ensemble. Its rank histogram is close to being uniform but not quite uniform; indeed, the latter was not to be expected. Sea level pressure is a synoptic variable with strong spatial correlation throughout the ensemble domain, and there were only 39 days in the evaluation period. The probability integral (PIT) histograms in Figure 6 accentuate the underdispersion in the raw ensemble and the bias-corrected ensemble.

3.3 Temperature forecasts

We now summarize the results for forecasts of surface temperature, a case of primary interest to the public (Murphy and Winkler 1979). The 2-m temperature forecasts were

Table 5: Comparison of predictive PDFs for sea level pressure over the Pacific Northwest. The bias-corrected ensemble and the EMOS forecasts were trained on a sliding 40-day period.

	$66\frac{2}{3}\%$ Prediction Interval		Score	
	Coverage	Average Width	CRPS	IGN
Climatological forecast	67.0	11.83	3.32	3.19
Raw ensemble	53.9	3.93	1.61	4.84
Bias-corrected ensemble	40.7	2.77	1.66	6.01
EMOS forecast	67.6	4.75	1.39	2.33

obtained as an average of the predicted lowest sigma level temperature and the predicted ground temperature. Similar to the sea level pressure forecasts, we used a sliding 40-day training period, and we considered the same region and the same test period. We omit the results for the climatological forecast which is even less competitive than for sea level pressure, given seasonal and topographic effects. The unit used for the temperature forecasts is degrees Kelvin.

Figure 7 shows how the estimates of the EMOS coefficients evolve over the test period. Figure 7(a) shows the estimated intercept which is consistently small and negative. Figures 7(b), (c), (d), (e), and (f) show the estimated EMOS weights, respectively. The weights for the AVN-MM5 forecast reached a maximum of 0.61 and were consistently the highest among the five ensemble member models. The weights for the ETA-MM5 forecast and for the NOGAPS-MM5 forecast were smaller but still substantial; those for the NGM-MM5 forecast were generally negligible or zero; and the weights for the GEM-MM5 forecast were initially negligible, before increasing to substantial levels. These results can be interpreted in terms of collinearity and ensemble member model skill. The correlation coefficient between the ETA-MM5 and the NGM-MM5 forecasts was the highest among the forecast pairs. To avoid collinearity, EMOS retained only one of them. The AVN-MM5 forecast was the most accurate member model and received the highest EMOS weights. Figures 7(f) and 7(g) show the estimated variance coefficients c and d, respectively. The estimates of d were consistently substantial.

23

(a) Raw Ensemble　　**(b) Bias–Corrected Ensemble**　　**(c) EMOS Ensemble**
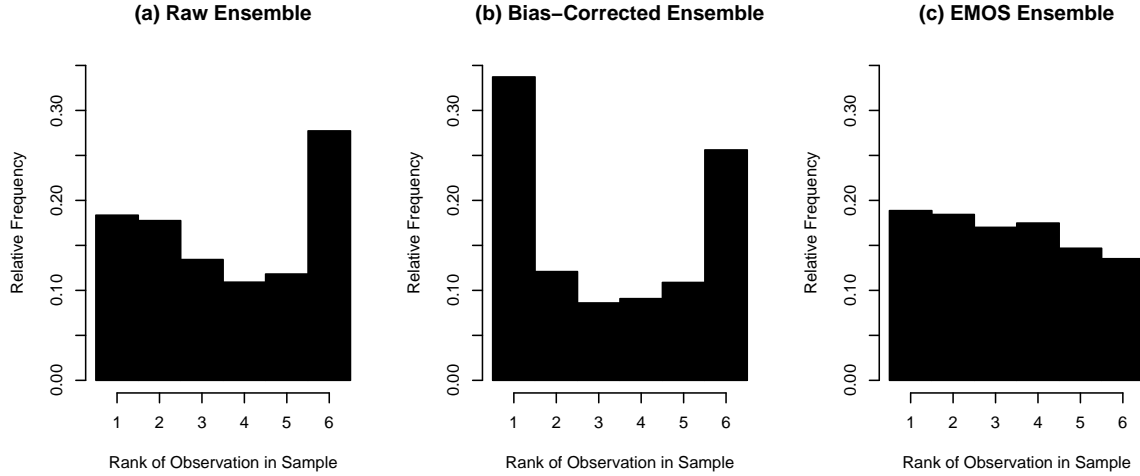
Figure 5: Verification rank histograms for ensemble forecasts of sea level pressure over the Pacific Northwest: (a) Raw ensemble. (b) Bias-corrected ensemble. (c) EMOS ensemble.

Table 6 confirms that the AVN-MM5 forecast was the most accurate among the ensemble member forecasts, both before and after bias correction. Bias correction resulted in percentage improvements in the RMSE of the ensemble member forecasts between 4% and 14%, and the NOGAPS-MM5 forecast showed the highest percentage improvement. The results in terms of the MAE were similar. The deterministic-style EMOS forecast was the most accurate, even though the percentage improvement over the bias-corrected ensemble was less pronounced than for forecasts of sea level pressure.

We now turn to a discussion of probabilistic forecast skill. Table 7 shows that the bias-corrected ensemble was slightly better calibrated than the raw ensemble. However, both the raw ensemble and the bias-corrected ensemble were strikingly underdispersive, and this was reflected in the CRPS and IGN scores, which were computed on the basis of standard ensemble smoothing. When computed directly from the ensemble CDF, the CRPS scores for the raw ensemble and for the bias-corrected ensemble were 2.13 and 1.95, respectively. The EMOS forecast performed best, with a CRPS score that was 15% lower than for the bias-corrected ensemble, and an IGN score that was 13 points lower. The verification rank histograms and PIT histograms are shown in Figures 8 and 9. The PIT histograms accentuate the underdispersion of the ensemble forecasts, while

24

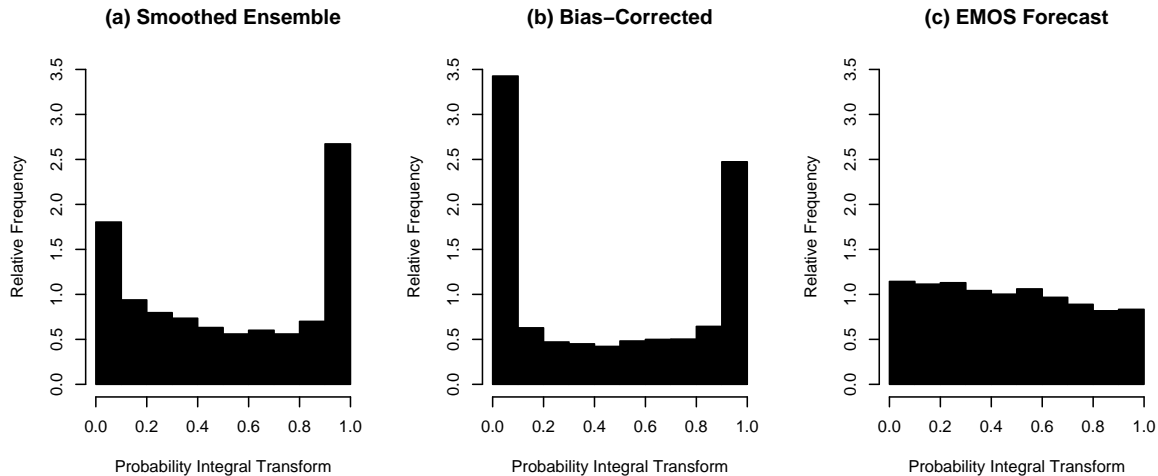**(a) Smoothed Ensemble**   **(b) Bias−Corrected**   **(c) EMOS Forecast**

Figure 6: Probability integral transform (PIT) histograms for PDF forecasts of sea level pressure over the Pacific Northwest: (a) Smoothed ensemble forecast. (b) Bias-corrected smoothed ensemble forecast. (c) EMOS forecast.

the histograms for the EMOS ensemble are close to being uniform.

4   Discussion

It is well documented in the literature that multiple regression or superensemble techniques improve the deterministic-style forecast accuracy of ensembles systems (Krishnamurti et al. 1999, 2000; Kharin and Zwiers 2002). Regression-based forecasts correct for model biases and therefore are more accurate than the ensemble mean forecast. The novelty of our ensemble model output statistics (EMOS) approach is three-fold. We constrain the regression coefficients to be nonnegative, thereby allowing for a more direct interpretation of the EMOS coefficients in terms of ensemble member model skill. EMOS identifies ensemble members whose relative contributions are negligible, typically as a result of collinearity. The method ignores those members when finding the EMOS predictive mean, an optimal bias-corrected weighted average of the ensemble member forecasts that provides a highly accurate deterministic-style forecast. For estimating the EMOS coefficients, we use the novel method of minimum CRPS estimation. Finally, we apply linear regression techniques to obtain full predictive PDFs, rather than
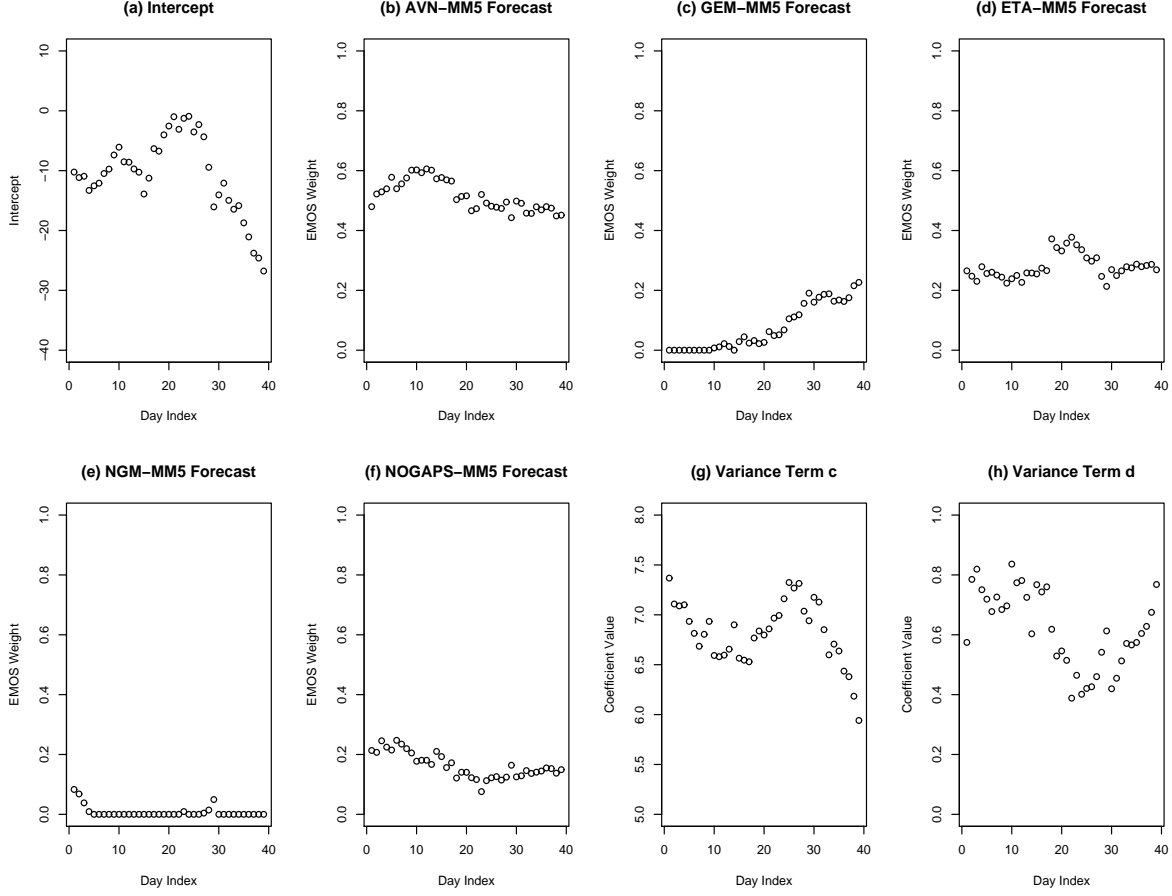
Figure 7: Coefficient estimates for EMOS forecasts of surface temperature over the Pacific Northwest. (a) Intercept. (b), (c), (d), (e), and (f): Member model weights. (g) and (h): Variance terms c and d.

deterministic-style forecasts, for continuous weather variables. The EMOS predictive PDFs are Gaussian, and they take account of the spread-skill relationship, in that the predictive variance is a linear function of the ensemble spread. However, EMOS adapts to the absence of spread-error correlation, by estimating the variance coefficient d as negligibly small. Monte Carlo simulation from the Gaussian predictive PDF is straightforward, and forecast ensembles of any size can be generated. An alternative, and likely preferable, way of forming an m-member ensemble from the predictive PDF is by taking the forecast percentiles at level $\frac{i}{m+1} \times 100\%$, for $i = 1, \ldots, m$.

We applied the EMOS technique to sea level pressure and surface temperature fore-

Table 6: Comparison of deterministic-style forecasts of surface temperature over the Pacific Northwest. The climatological, bias-corrected, and EMOS forecasts were trained on a sliding 40-day period.

	MAE	RMSE
AVN-MM5	2.45	3.15
GEM-MM5	2.64	3.40
ETA-MM5	2.52	3.23
NGM-MM5	2.56	3.28
NOGAPS-MM5	2.96	3.76
AVN-MM5 bias-corrected	2.31	3.00
GEM-MM5 bias-corrected	2.48	3.24
ETA-MM5 bias-corrected	2.39	3.10
NGM-MM5 bias-corrected	2.42	3.13
NOGAPS-MM5 bias-corrected	2.50	3.25
Mean of raw ensemble	2.49	3.18
Mean of bias-corrected ensemble	2.28	2.95
EMOS forecast	2.23	2.91

casts over the North American Pacific Northwest in Spring 2000, using the University of Washington mesoscale ensemble (Grimit and Mass 2002). The EMOS predictions were more accurate when compared to the member model forecasts, bias-corrected member model forecasts, the ensemble mean forecast, and the ensemble mean of the bias-corrected member models. We also assessed the probabilistic forecast skill of the EMOS predictive PDFs. When compared to the bias-corrected ensemble, EMOS PDF forecasts of sea level pressure had substantially better CRPS and IGN scores. The EMOS PDFs were much better calibrated than the raw ensemble or the bias-corrected ensemble, and they were sharp, in that EMOS prediction intervals were much shorter on average than prediction intervals based on climatology. Perhaps surprisingly, the EMOS forecasts of sea level pressure were frequently sharper than the raw ensemble forecasts. With small modifications, as explained in Section 2.3, EMOS applies to all ensemble systems, including weather and climate, synoptic-scale, poor person's, multi-model, multi-analysis, perturbed observations, singular vector, and bred ensembles. EMOS can be applied to gridded ensemble output, thereby providing probabilistic forecasts on a grid. The re-

Table 7: Comparison of predictive PDFs for surface temperature over the Pacific Northwest. The bias-corrected ensemble and the EMOS forecasts were trained on a sliding 40-day period.

	$66\frac{2}{3}\%$ Prediction Interval		Score	
	Coverage	Average Width	CRPS	IGN
Raw ensemble	28.7	2.55	2.07	21.45
Bias-corrected ensemble	31.1	2.44	1.89	15.50
EMOS forecast	68.6	5.41	1.61	2.49

sulting forecast fields can be visualized in the form of percentile maps, as in Figure 5 of Raftery et al. (2003). In our experiment, we used observations to estimate the EMOS coefficients, but this could also be done using an analysis.

Bias correction results in more accurate deterministic-style forecasts, and bias correction reduces ensemble spread, by pulling the individual member model forecasts towards the verification mean (Eckel 2003). Verification rank histograms typically become more symmetric after bias correction, as in our Figure 8, or in Figure 46 of Eckel (2003). However, bias correction does not result in improved calibration, and the need for statistical post-processing remains. We anticipate significant improvements in probabilistic forecast skill through the use of advanced bias correction schemes, followed by statistical post-processing of the bias-corrected member model ensemble. Further research in this direction is desirable.

We close with a discussion of potential extensions as well as limitations of the EMOS technique. The predictive PDFs produced by the EMOS method are Gaussian and therefore unimodal. This is unlikely to be a disadvantage for a five-member ensemble, such as the University of Washington ensemble that we considered. However, larger ensembles occasionally suggest multimodal forecast PDFs. The ensemble smoothing approach of Wilks (2002) and the Bayesian model averaging approach of Raftery et al. (2003) address this issue.

We obtained EMOS forecasts of sea level pressure and surface temperature. These are variables for which the forecast error distributions are approximately Gaussian. The

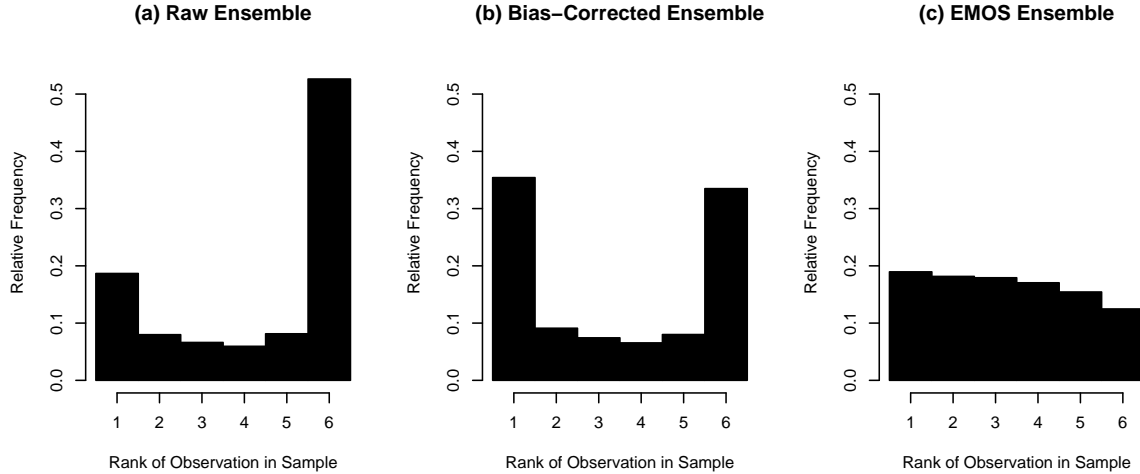**(a) Raw Ensemble** **(b) Bias–Corrected Ensemble** **(c) EMOS Ensemble**

Figure 8: Verification rank histograms for ensemble forecasts of surface temperature over the Pacific Northwest: (a) Raw ensemble. (b) Bias-corrected ensemble. (c) EMOS ensemble.

forecast error distributions for other variables, such as precipitation or cloud-cover, are unlikely to be close to normal. Wilks (2002) proposes ways of transforming forecast ensembles to Gaussian distributions, and EMOS can be applied to the transformed ensemble. Another approach uses the framework of generalized linear models (McCullagh and Nelder 1989), and this remains to be explored.

Our method produces predictive PDFs of continuous weather variables at a given location, but it does not reproduce the spatial correlation patterns of observed weather fields. Gel et al. (2004) suggest a way of creating ensembles of entire weather fields, each of which honors the spatial correlation structure of verifying fields. However, this approach uses only one numerical weather prediction model rather than an ensemble of forecasts. This method could be combined with EMOS to yield calibrated ensembles of entire weather fields, by simulating correlated error fields and adding them to the spatially varying predictive mean of the EMOS forecasts. Such an approach could also be viewed as a dressing method (Roulston and Smith 2003). A different, somewhat simplistic idea is what might be called a probability field ensemble, that is, an ensemble of m, say $m = 5$, weather fields showing the percentiles of the EMOS predictive distribution function at the levels $\frac{i}{m+1} \times 100\%$ for $i = 1, \ldots, m$, respectively. Probability field
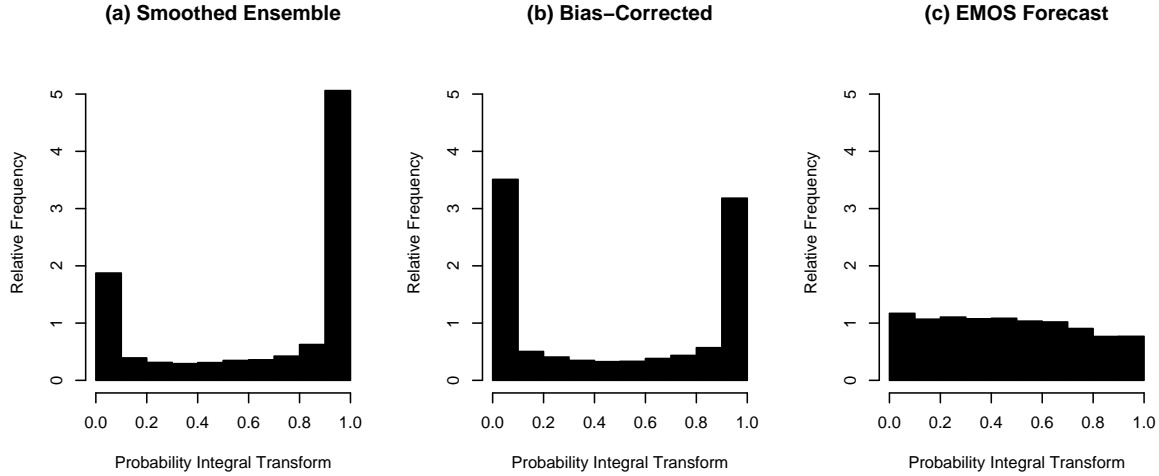
Figure 9: Probability integral transform (PIT) histograms for PDF forecasts of surface temperature over the Pacific Northwest: (a) Smoothed ensemble forecast. (b) Bias-corrected smoothed ensemble forecast. (c) EMOS forecast.

ensembles do not reproduce the spatial correlation structure of observed weather fields, nor do they take account of dynamical features. However, a probability field ensemble could be interpreted as a sample of equally likely weather fields, with respect to any fixed location, which may facilitate the interpretation, and may foster the acceptance and the use of probabilistic forecasts.

Acknowledgements

References

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523.

Birgé, L., and P. Massart, 1993: Rates of convergence for minimum contrast estimators. *Probab. Theory Rel. Fields*, **97**, 113–150.

Dawid, A. P., 1984: Statistical theory: The prequential approach. *J. Roy. Stat. Soc. Ser. A*, **147**, 278–292.

Déqué, M., J. T. Royer, and R. Stroe, 1994: Formulation of gaussian probability forecasts based on model extended-range integrations. *Tellus*, **A46**, 52–65.

Eckel, F. A., 2003: Effective mesoscale, short-range ensemble forecasting. Ph.D. Dissertation, Department of Atmospheric Sciences, University of Washington, Seattle, Washington. [Available online at www.atmos.washington.edu/~ens/pubs_n_pres. html]

——, and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132–1147.

Gel, Y., A. E. Raftery, and T. Gneiting, 2004: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion and rejoinder). *J. Amer. Stat. Assoc.*, in press.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

Good, I. J., 1952: Rational decisions. *J. Roy. Statist. Soc., Ser. B*, **14**, 107–114.

Gneiting, T., and A. E. Raftery, 2004: Strictly proper scoring rules for probabilistic forecasts. Manuscript in preparation.

——, ——, F. Balabdaoui, and A. Westveld, 2003: Verifying probabilistic forecasts: Calibration and sharpness. *Proc. Workshop on Ensemble Forecasting*, Val-Morin, Québec. [Available online at www.cdc.noaa.gov/~hamill/ef_workshop_2003.html]

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205.

——, and ——, 2004: Forecasting mesoscale uncertainty: Short-range ensemble forecast error predictability. *Preprints, 16th Conf. on Numerical Weather Prediction*, Seattle, Washington, paper 24.3. [Available online at www.atmos.washington.edu/~ens/pubs_n_pres.html]

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

——, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

——, and ——, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Huber, P. J., 1964: Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101.

——, 1981: *Robust Statistics*. John Wiley, 308 pp.

Jewson, S., A. Brix, A., and C. Ziehmann, 2003: A new framework for the assessment and calibration of medium range ensemble temperature forecasts. Preprint. [Available online at www.arXiv.org/abs/physics/0308057]

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate*, **15**, 793–799.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendan, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550.

——, ——, Z. Zhang, T. E. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendan, 2000: Multimodel ensemble forecasts for weather and seasonal cli-

mate. *J. Climate*, **13**, 4196–4216.

Mass, C. F., 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting*, **18**, 75–79.

——, M. Albright, D. Ovens, R. Steed, M. MacIver, E. Grimit, T. Eckel, B. Lamb, J. Vaughan, K. Westrick, P. Storck, B. Colman, C. Hill, N. Maykut, M. Gilroy, S. A. Ferguson, J. Yetter, J. M. Sierchio, C. Bowman, R. Stender, R. Wilson, and W. Brown, 2003: Regional environmental prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **84**, 1353–1366.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models*, 2nd ed. Chapman & Hall, 511 pp.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

Murphy, A. H., and R. L. Winkler, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc*, **60**, 12–19.

Pfanzagl, J., 1969: On the measurability and consistency of minimum contrast estimates. *Metrika*, **14**, 249–272.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1993: *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, 963 pp.

Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2003: Using Bayesian model averaging to calibrate forecast ensembles. Technical Report no. 440, Department of Statistics, University of Washington. [Available online at www.stat. washington.edu/tech.reports]

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of sample size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489.

Rosenblatt, M., 1952: Remarks on a multivariate transformation. *Ann. Math. Stat.*, **23**, 470–472.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

——, and ——, 2003: Combining dynamical and statistical ensembles. *Tellus*, **A55**, 16–30.

Stefanova, L., and T. N. Krishnamurti, 2002: Interpretation of seasonal climate forecast using Brier score, the Florida State University superensemble, and the AMIP-I dataset. *J. Climate*, **15**, 537–544.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems, *Proc. Workshop on Predictability*, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts, *Forecast Verification. A Practicioner's Guide in Atmospheric Science*, Jolliffe, I. T., and D. B. Stephenson, Eds., 137–163.

Unger, D. A., 1985: A method to estimate the continuous ranked probability score. *Preprints, 9th Conf. on Probability and Statistics in Atmospheric Sciences*, Virginia Beach, Virginia, 206–213.

van den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6-10-day forecast. *Wea. Forecasting*, **3**, 457–465.

Weigend, A. S., and S. Shi, 2000: Predicting daily probability distributions of S&P500 returns. *J. Forecasting*, **19**, 375–392.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.*, **128**, 2821–2836.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127**, 956–970.